Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 151 (2019) 1201–1206

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Web Search and Data Mining (WSDM)
April 29 - May 2, 2019, Leuven, Belgium

Improvements for Determining the Number of Clusters in k-Means for Innovation Databases in SMEs

Amelec Viloria[a]*, Omar Bonerge Pineda Lezama[b]

[a] *Universidad de la Costa (CUC), Barranquilla 080003, Colombia*
[b] *Universidad Tecnológica Centroamericana (UNITEC), Tegucigalpa 11101, Honduras*

Abstract

The Automatic Clustering using Differential Evolution (ACDE) is one of the grouping methods capable of automatically determining the number of the cluster. However, ACDE continues making use of the strategy manual to determine the activation threshold of k, which affects its performance. In this study, the problem of ACDE is enhanced using the U Control Chart (UCC). The performance of the proposed method was tested using five data sets from the National Administrative Department of Statistics (DANE - Departamento Administrativo Nacional de Estadísticas) and the Ministry of Commerce, Industry, and Tourism of Colombia for the innovative capacity of Small and Medium-sized Enterprises (SMEs) and were assessed by the Davies Bouldin Index (DBI) and the Cosine Similarity (CS) measure. The results show that the proposed method yields excellent performance compared to prior researches for most datasets with optimal cluster number yet lowest DBI and CS measure. It can be concluded that the UCC method is able to determine k activation threshold in ACDE that caused effective determination of the cluster number for k-means clustering.

Keywords: k-means; automatic clustering; differential evolution; k activation threshold; U-Control Chart; SMEs.

* Corresponding author. Tel.: +573046238313
E-mail address: aviloria7@cuc.edu.co

10.1016/j.procs.2019.04.172

1. Introduction

In Colombia, according to Amelec, V. (2015) [1], SMEs play an important role in economic activity, which promotes the growth and development of innovative ideas that facilitate the processes of entrepreneurship. The current perspective of Colombia SMEs establishes certain challenges that must be taken into account to start new strategies that are aimed at continuous improvement and sustainable growth of these companies. Some of the challenges that currently face the SMEs are: the strengthening of their human resources and organizational culture the transfer of knowledge the improvements in quality and productivity, the creation of business networks (clusters), the development of R&D&I activities that contribute to the generation of innovation, expansion in the domestic and international market, and optimization of their processes (Lis-Gutiérrez M. et al; 2018) [2].

The measurement of innovation has been a topic with great interest in recent times. There are many studies that have been carried out, due to the need of identifying the behavior of the innovative process of the companies (Bartels and Koria, 2014) [3]. In Colombia, most of these studies have been carried out following the Oslo Manual approach from the Organization for Economic Cooperation and Development (OECD). The results obtained in these studies allow governments to design public policies consistent with the actual situation in the field of innovation and technology development. In the country, most of the researches based on innovation surveys (DANE, 2017) [4], have focused on analyzing the data obtained up to the middle of the last decade. However, the absence of academic works that deal with the most recent information through the application of statistical techniques of inferential cut is notorious.

In this sense, the present study seeks to show those associations that allow to identify the behavior patterns in innovation processes in Small and Medium-sized Enterprises of Colombia, based on the Technological Innovation Survey (EDIT - Encuesta de Innovación Tecnológica, 2016-2017) [4], developed by the DANE, which includes aspects such as investment in science, technology and innovation activities, the human capital related to these activities, the relationship with other actors of the innovation system, the intellectual property, the quality certifications, the technical standards and regulations, and the impact on innovation.

Given the above, it is clear that there is a lot of information needed to process and identify behavior patterns in the SMEs that will allow the redefinition of policies in the field of innovation. For this purpose, the use of data mining techniques is propitious. However, finding the methods or integrating them to obtain the best yields when determining the cluster in the SMEs sector is necessary. The Automatic Clustering using Differential Evolution (ACDE) is one of the grouping methods capable of automatically determining the number of clusters. Nevertheless, ACDE continues using the strategy manual to determine the activation threshold of k, which affects its performance. In this study, the problem of ACDE is enhanced using the U Control Chart (UCC).

2. Data and Proposed Method

The information used corresponds to the Technological Innovation Survey development, which was applied by the Administrative Department of Statistics (DANE) to the manufacturing companies of the country, for the period 2016-2017, given the high number of observations made, the availability of the microdata, and their quality. The study covers a total of 6,954 manufacturing SMEs, from all sectors of the economy, located throughout the Colombian territory (DANE, 2017) [4]. The survey form is presented in several chapters, each related to a different aspect, namely the impact of innovation, investment in science, technology and innovation activities, the human capital involved in those activities, the relationship with innovation systems, and issues related to intellectual property, quality certifications, and technical standards and regulations.

Regarding the database debugging, those variables that did not contribute to the analysis because their response modalities were very different were excluded (Jolliffe, 2002) [5]. For example, questions whose answer is YES or NO, where 98% answered YES or NO were discarded. This information biases the analysis of the set of variables.

It is also important to note that in the variable called typology, the "potentially innovative" response modalities were combined with "strictly innovative" ones, since their individual contribution to the analysis is low and the fact that each one is alone, skews the obtained results.

Subsequently, a multiple correspondence analysis (MCA) was carried out for each chapter of the survey with a specific topic and some typologies were determined by group classification. This method consists of making associations between the variables studied in the study units in such a way that they are grouped into some factors, according to their similarities (Chakraborty and Das, 2018) [6]. Finally, the analysis of results was carried out, seeking to find associations that would allow to characterize the innovation processes among the study units, treated in each of the variables. The final base used is shown in Table 1.

Table 1. Database on innovation in SMEs

Datasets	Number of records	Attributes	Class (k)
Base 1: Innovation and impact in the company	300	2	3
Base 2: Investment in scientific, technological and innovation activities (STi)	150	4	3
Base 3: Personnel engaged in STi activities	214	9	6
Base 4: Relations with actors of the national STi system	178	13	3
Base 5: Intellectual property, quality certifications, technical and regulatory standards	990	13	11

The k-means method is one of the hard partition methods in cluster analysis of the data mining field. The k-means has advantages, i.e. it is easy to implement, grouping a large dataset, and with stable performance over different problems (Chakraborty and Das, 2018) [6]. However, the clustering results of k-means depend on a certain number of clusters as inputs, if the estimated number of clusters does not tally with the final solution, the chances of clustering are very low. Meanwhile, getting the number of k as an input on k-means is still not an easy task because the user requires prior specification number of the cluster. This condition is termed a local optimum problem. In practice, the local optimum problem is overcome by applying the method several times with a different number of k, then choosing the best results. Automatic clustering methods are one solution that helps the user determine the optimal number of clusters (Garcia and Flores, 2016) [7]. Therefore, the automatic clustering method is an effective solution for this problem.

Researches on the determination of the number of clusters used automatic clustering methods which are based on the Evolutionary Computation (EC) technique. The K-means method has done a lot and has been published with different methods, namely Automatic Clustering using Differential Evolution (ACDE) (Das et al., 2008) [8], combination methods between PSO and k-means on Dynamic Clustering with Particle Swarm Optimization (DCPSO) and Genetic Clustering for unknown k clustering (GCUK).

Automatic clustering methods have been used to determine the number of clusters in the k- means but are yet to achieve an accurate cluster result. Therefore, it is necessary to improve the performance of automated grouping methods used for determining the number of clusters. The ACDE method is the most popular EC technique which has effectively improved the performance of automatic clustering methods proposed by previous researchers (Das et al., 2008) [8]. ACDE predicated on the Differential Evolution (DE) method is one of the strongest, fastest, and most efficient global search heuristics methods in the world that is very easy to use with high-dimensional data. It can be employed using polynomial functions and other functions because it is easy to change the values of control variables such as NP, F, and CR to obtain good search results (Ramadas et al., 2016) [9].

The ACDE was then developed by (Kuo et al., 2013) [10] and the combination of ACDE and k-means methods was termed the automatic clustering approach based on the differential evolution method combined with k-means for crisp clustering method aimed at improving clustering performance in the k-means method (ACDE-k-means).

The ACDE method is capable of finding the number of clusters automatically and is able to balance the evolutionary process of DE methods to achieve better partitions than the classic DE. However, the DE classic method still depends on user's considerations to determine the k activation threshold thereby affecting the performance of the DE method.

The U-Control Chart (UCC) method is employed to determine the k activation threshold that is used for the initial step to get the value of the variables sought before initialization of the variable vector. The UUC is a method from statistical process control (SPC) which has proved to be effective in solving the problem of management control attributes (Kaya, 2009) [11], other methods such us P-Control Chart and C-Control Chart are also methods but not used. This research focuses just on UCC. The UCC used to average the data to be measured is then reduced and added to find upper and lower bound values on the number of attributes for the searched variables. A product is said to have a good quality if the average value is at a threshold or the average value is between the upper and lower bound. Based on the above assumption, the data is good if it is within the threshold of the U-Control Chart.

The aim of the cluster validity index is to measure how efficient cohesion and separation are (Garcia and Flores, 2016) [7]. Compactness is used to measure variation or pattern of data within a cluster, and separation shows cluster isolation from each other using matrix distance (usually their used Euclidean distance). There are many indices the validity index cluster can use, but in this study, only Davies Bouldin Index (DBI) / Cosine Similarity measure (CS) is used as cluster validity index to help find the right number of clusters because it has been widely used and is state-of-the-art. In the study, the same cluster validity (DBI /CS) of the original ACDE method is used as fitness function (Tam et al., 2017) [12].

In this research, a combination of the U-Control Chart (UCC) and Automatic Clustering using Differential Evolution method is proposed to determine the number of clusters on k-means. The aim of the UCC method is to control k activation threshold of the Automatic Clustering using Differential Evolution method. The latter will automatically search the optimal number of clusters in the data as required by k-means. The representation of chromosome used is based on (Das et al., 2008) [8]. Because the Automatic Clustering using Differential Evolution method produces premature clusters, the k-means is implemented to repair these events.

4. Results and Discussions

The experiments were conducted using a computing platform with Intel Celeron 2.16 GHz CPU, 8 GB RAM and Microsoft Windows 10 Home 64-bit used as the operating system, and MATLAB version R2016a used as the data analytics tool (Kamatkar S. et al; 2018) [13]. MATLAB would produce a model performance as the calculation output, such as average value best cluster DBI and CS measure. The proposed method was tested using the Innovation Database in SMEs. Parameter setting for proposed method based on the recommendation of Das et al. (2008) [8] is as follows: maxiter = 200, pop-size=10*dim, CRmax = 1.0 and CRmin = 0.5. Max-iteration indicates the amount of iteration, pop-size is the size of the population, cross-over probability is used to initialize the position of a particle or chromosome.

First, an experiment was conducted on 5 datasets from UCI by using only ACDE k-means without the UCC method. Classes on the data were omitted to analyze optimal partition in a data. In the second experiment, the U-Control Chart (UCC) method was implemented to solve the problem of k activation threshold automatically without requiring the user to enter the required values in ACDE-k-means for determining the number of clusters of k-means, whereas, k-means was implemented to do repair grouping (Varela Izquierdo N. et al; 2018) [14], (Gaitán-Angulo M. et al; 2018) [15]. A more detailed comparison of the first and the second experiment is presented in Table 2. The best model automatic clustering on each dataset is highlighted with boldfaced print and the best optimal cluster result is marked with (1) and (2) squared on each dataset. As shown in Table 2, the second experiment UCC+ACDE-k-means is outperforming in almost all datasets with respect to DBI. The results obtained were applied (UCC+ACDE-k-means) on the basis of data regarding innovation in SMEs. The results are shown in Table 3.

Table 2. Results comparison ACDE-k-means only vs UCC+ACDE-k-means.

Datasets	Class optimal (k)	DBI		k		CS		k	
		(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Base 1	3	0.5642	0.4830	3*	3*	0.6695	0.5495	3*	3*
Base 2	3	0.0390	0.5090	3*	3*	0.0425	0.3043	3*	3*
Base 3	6	0.6568	0.4989	3#	8#	0.0480	0.0358	6*	6*
Base 4	3	0.3956	0.2083	3*	3*	0.0071	0.2677	3*	3*
Base 5	5	1.2728	1.0324	5*	4#	0.7818	0.9209	2#	3#

(1)ACDE-$_k$-means only (2)UCC-ACDE-$_k$-means ; *Number cluster optimal #not optimal

Table 3. Distribution of the clusters according to the dimensions of the Colombian Technological Innovation Survey (EDIT)

Aspects of the Innovation Survey	**Composition of the Cluster**		
Innovation and impact in the company	A - Innovators - 33.26%	B - Beginners - 13.75%	C - Laggards - 52.99%
Investment in scientific, technological and innovation activities (STi)	A - Innovators - 11.88%	B - Laggards - 88.12%	
Personnel engaged in STi activities	A - Innovators - 11.49%	B - Followers - 13.49%	C - Laggards - 75.09%
Relations with actors of the national STi system	A - A lot of research and relationship - 5.89%	B - Incipient research and relationship - 19.26%	C - Little research and relationship -74.91%
Intellectual property, quality certifications, technical and regulatory standards	A - Innovators - 7.81%	B - Followers - 24.26%	C - Laggards - 67.93%

Table 3 shows the distribution of the three groups. The first of them, called innovators (A), belongs to 33.26% of the companies, which are characterized by having high levels in terms of new organizational methods and production, and original and improved goods for the national market. In addition, no obstacles were found in terms of lack of resources, lack of qualified personnel, difficulty in complying with information on the market, compliance with regulations and technical regulations, among others. In the second one, 13.75% of SMEs belong to the group of beginners (B), which are characterized because they found some obstacles in their innovative process, such as: uncertainty regarding the technical execution of projects and facing the demand for goods and innovative services, low profitability in innovation, limited qualified personnel, capacity of imitation by third parties, among others (Table 1). In the third one, 52.99% of SMEs belong to the group of laggards (C), characterized by finding obstacles that are difficult to overcome in their innovative process, such as uncertainty regarding the demand for innovative goods and services and against the technical execution of projects, the low possibility of cooperation with other companies, the ease of imitation by third parties, the difficulty to access financing, and the lack of qualified personnel, among others.

5. Conclusions

The use of U-Control Chart (UCC) method with Automatic Clustering using Differential Evolution (ACDE) to determine the number of clusters in k-means has proven to increase the performance of ACDE. As noted in the previous analysis, when examining the variables evaluated in the innovation processes of SMEs, some important findings were obtained, such as that companies can be grouped into three clusters: The first one, made up by successful companies, which obtained, in almost all evaluated issues, outstanding qualifications and, although they make up a relatively small group, they are successful in their innovation processes. The second group can be characterized as stable or indifferent companies, which recorded average levels of innovative performance based on the various variable categories analyzed. Finally, the third group consisting of approximately three quarters of the

companies, which presented a low average performance in the various issues evaluated. This is the group of lagging non-innovative SMEs, distinguished by their low capacity for relationship, lack of qualified personnel, low investment and dedication to activities related to innovation and, therefore, a low level in the development of new products and services for the market. This is consistent with the Colombian reality, considering that Colombia is a country with a low level of investment in science, technology and innovation activities.

Further research may be added to other control charts derived from Statistical Process Control (SPC) such as P-Control Chart (PCC) and C-Control Chart (CCC). According to (Kaya, 2009) [11], the SPC method can easily detect changes in the data of a process that may affect the quality of the results.

References

[1] Amelec, V. (2015). Increased efficiency in a company of development of technological solutions in the areas commercial and of consultancy. Advanced Science Letters, 21(5), 1406-1408.

[2] Lis-Gutiérrez M., Gaitán-Angulo M., Balaguera MI., Viloria A., Santander-Abril JE. (2018) Use of the Industrial Property System for New Creations in Colombia: A Departmental Analysis (2000–2016). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[3] Bartels, F.; Koria, R. 2014. Mapping, measuring and managing African national systems of innovation for policy and development: the case of the Ghana national system of innovation. African J. Science, Technol., Innov. Developm. 6(5):383-400.

[4] DANE. 2017. Documento metodológico encuesta de desarrollo e innovación tecnológica en la industria Manufacturera. Bogotá: DANE. 43p.

[5] Jolliffe, I. 2002. Principal component analysis. Hoboken: John Wiley & Sons, 488p.

[6] Chakraborty, S., Das, S., 2018. Simultaneous variable weighting and determining the number of clusters—A weighted Gaussian means algorithm. Stat. Probab. Lett. 137, 148– 156. https://doi.org/10.1016/j.spl.2018.01.015

[7] Garcia, A.J., Flores, W.G., 2016. Automatic Clustering Using Nature-Inspired Metaheuristics: A Survey. Appl. Soft Comput. https://doi.org/10.1016/j.asoc.2015.12.001

[8] Das, S., Abraham, A., Konar, A., 2008. Automatic Clustering Using an Improved Differential Evolution Algorithm. IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans 38, 218–237. https://doi.org/10.1109/TSMCA.2007.909595

[9] Ramadas, M., Abraham, A., Kumar, S., 2016. FSDE-Forced Strategy Differential Evolution used for data clustering. J. King Saud Univ. - Comput. Inf. Sci. https://doi.org/10.1016/j.jksuci.2016.12.005.

[10] Kuo, R.., Suryani Erma, Yasid, A., 2013. Automatic Clustering Combining Differential Evolution Algorithm and k-Means Algorithm. Proc. Inst. Ind. Eng. Asian Conf. 2013 1207–1215. https://doi.org/10.1007/978-981-4451-98-7

[11] Kaya, I., 2009. A genetic algorithm approach to determine the sample size for attribute control charts. Inf. Sci. (Ny). 179, 1552–1566. https://doi.org/10.1016/j.ins.2008.09.024

[12] Tam, H., Ng, S., Lui, A.K., Leung, M., 2017. Improved Activation Schema on Automatic Clustering Using Differential Evolution Algorithm. IEEE Congr. Evol. Comput. 1749–1756. https://doi.org/10.1109/CEC.2017.7969513

[13] Kamatkar S.J., Tayade A., Viloria A., Hernández-Chacín A. (2018) Application of Classification Technique of Data Mining for Employee Management System. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[14] Varela Izquierdo N., Cabrera H.R., Lopez Carvajal G., Viloria A., Gaitán Angulo M., Henry MA. (2018) Methodology for the Reduction and Integration of Data in the Performance Measurement of Industries Cement Plants. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham.

[15] Gaitán-Angulo M. Jairo Enrique Santander Abril, Amelec Viloria, Julio Mojica Herazo, Pedro Hernández Malpica, Jairo Luis Martínez Ventura, Lissette Hernández-Fernández. (2018). Company Family, Innovation and Colombian Graphic Industry: A Bayesian Estimation of a Logistical Model. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham.